De quem é essa assinatura?

Cientistas da computação criam novas estratégias para desfazer ambiguidades em referências bibliográficas

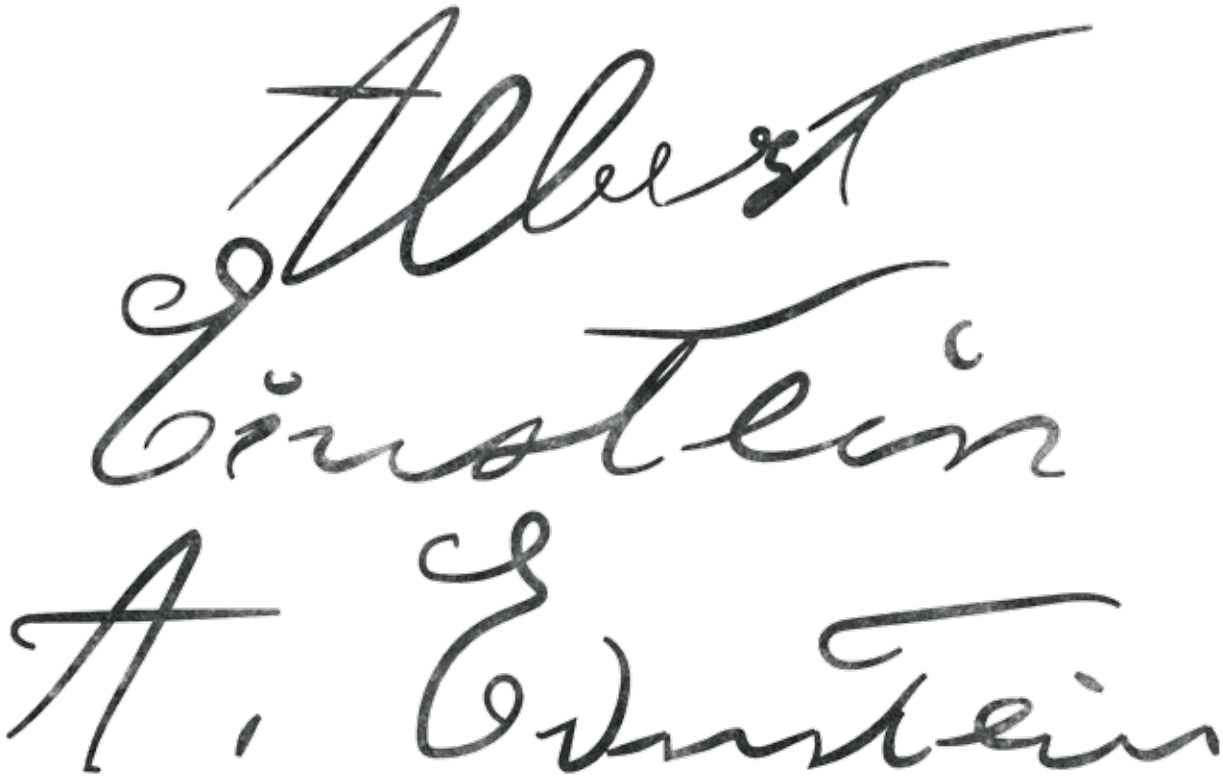
Cientistas da computação da Universidade Federal do Rio de Janeiro (UFRJ) propuseram uma nova abordagem para enfrentar o problema da ambiguidade de assinaturas de autores científicos em referências bibliográficas, que faz com que a produção de um pesquisador ora seja confundida com a de colegas que adotam abreviação idêntica, ora seja difícil de agrupar e avaliar, porque o mesmo pesquisador utiliza assinaturas diferentes. Em um artigo publicado em maio na revista *Scientometrics*, a cientista da computação Janaina Gomide e seu orientador de doutorado Daniel Ratton Figueiredo, professor do Programa de Engenharia de Sistemas e Computação da UFRJ, mostraram a existência de comportamentos que se repetem entre os autores que usam várias assinaturas.

Um deles é a mudança rara ou acidental da assinatura em algum dos *papers* publicados, uma espécie de ponto fora da curva causado por um erro ou descuido do autor ou da revista. Outro padrão é o do pesquisador que assina de uma maneira no começo da carreira e, a partir de certo momento, passa a assinar de outra forma, caso, por exemplo, de mulheres que mudam de sobrenome quando casam ou se separam. E, por fim, há um padrão mais difícil de detectar, o do pesquisador que assina de várias formas sem se preocupar com uma normatização de sua assinatura.

Os pesquisadores avaliaram a incidência desses comportamentos em dois ambientes distintos. Um foi a base de dados do Digital Bibliographic Library Project (DBLP), que reúne a produção de cientistas da computação e é usada com frequência como referência em estudos sobre ambiguidade, porque já foram mapeados os casos em que há padrões de assinatura repetitivos. Também foram avaliados 881 pesquisadores brasileiros cujos perfis no Google Scholar exibiam mais de um tipo de assinatura, selecionados entre os bolsistas de produtividade do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Revelou-se que a substituição acidental de assinatura é a mais frequente, com 43% dos registros no DBLP e 53% do Google Scholar. A troca num determinado momento da carreira foi responsável por um terço dos casos no DBLP e 18% do Google Scholar. Já a oscilação frequente de assinaturas se revelou um pouco mais comum entre os autores nacionais do Google Scholar, com um terço dos casos, e menos frequente no DBLP, que reúne pesquisadores de vários países, com 25% do total. Uma explicação para a mudança frequente de abreviação entre brasileiros é que o uso no país de nomes compostos e mais de um sobrenome favorece a confusão. “Temos muitos sobrenomes e os utilizamos de forma livre, enquanto os autores dos Estados Unidos são identificados, em geral, apenas pelo primeiro e o último nome”, esclarece Daniel Figueiredo, ele próprio uma vítima do problema: a maioria de seus artigos científicos leva a assinatura Figueiredo, D. R., mas há outros com variantes como Figueiredo, Daniel ou Figueiredo, Daniel R.

O passo seguinte do trabalho foi avaliar as redes de colaboração dos pesquisadores que publicam com mais de uma assinatura. Observou-se que cada uma



das três classes – o uso ocasional de uma outra assinatura, a troca de assinatura num certo momento da carreira e o uso frequente de mais de uma assinatura – apresenta redes de colaboração com padrões claros e específicos, cujos perfis podem ser úteis para formular no futuro algoritmos capazes de ajudar a identificar nomes ambíguos. "Desvendar a ambiguidade de nomes é um problema clássico da computação e o que se tenta fazer sempre é encontrar todos os rótulos, ou tipos de assinaturas, que remetem a uma mesma pessoa", diz Janaina Gomide. "A contribuição do nosso trabalho foi mostrar as causas comuns da ambiguidade, que já eram conhecidas de forma intuitiva, mas ainda não haviam sido medidas, e propor sua utilidade na construção de novos algoritmos", completa Figueiredo.

CONFUSÃO NA AVALIAÇÃO

O interesse dos pesquisadores por esse tema se explica tanto pelo desafio de criar ferramentas computacionais para resolver um problema concreto quanto pela confusão que as ambiguidades causam na hora de medir a produção de um cientista, causando prejuízos em processos de avaliação ou em estudos bibliométricos que necessitam de informações precisas sobre autores. Em um levantamento publicado em 2012 no *Sigmod Record*, publicação trimestral da Association for Computing Machinery (ACM), o brasileiro Alberto Laender, professor do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais (UFMG), contabilizou 17 métodos computacionais distintos para resolver o problema da ambiguidade que eram utilizados na época. "Hoje, já deve haver pelo menos uns 30 algoritmos diferentes em uso", conta.

O grupo da UFMG elaborou três desses algoritmos. Um deles, conhecido como HHC (Heuristic-based Hierarchical Clustering), foi apresentado em 2007 e passou a ser usado pela DBLP, a mesma base de dados usada no estudo de Daniel Figueiredo, como uma das ferramentas mais simples para enfrentar o problema. Fruto de uma dissertação de mestrado defendida por Ricardo Cota na UFMG, o HHC reúne as informações bibliográficas vinculadas a uma assinatura e analisa

ILUSTRAÇÕES DANIEL KONDO

se há coautores que se repetem. Quando existe coincidência, avalia também se os títulos dos artigos têm palavras em comum ou se os autores participaram dos mesmos eventos científicos. A eficiência para desfazer a ambiguidade chegou perto de 80%. "O método passou a ser usado por sua simplicidade, mas a busca por algoritmos cada vez mais precisos continuou", diz Laender. "Há situações em que não há algoritmo capaz de resolver o problema. Entre autores da China, que têm sobrenomes frequentes e uma grande quantidade de abreviações coincidentes, chega a ser inviável."

Um segundo método criado por pesquisadores da UFMG foi o Sand (Self-training Associative Name Disambiguator), que agrupa referências bibliográficas de acordo com características comuns, como a presença de coautores, título e ano de publicação. Utilizando técnicas de inteligência artificial, consegue detectar, em sua etapa final, se há autores que, dadas as suas características, deveriam pertencer a determinados agrupamentos – e calcular as chances de que tais registros sejam referências ambíguas de outros autores já existentes. "Essas técnicas de classificação são bastante conhecidas e um dos nossos ex-alunos de doutorado, Anderson Ferreira, hoje professor da Universidade Federal de Ouro Preto, adaptou-as para a desambiguação. O Sand julga em diferentes classes as referências até chegar à conclusão de que um determinado autor tem de estar naquela classe", afirma Laender. E o terceiro método é o IDNi (Incremental Unsupervised Name Disambiguation), que associa diversas técnicas e é usado para avaliar novos trabalhos científicos incorporados a bases de dados, associando-os de forma automática a perfis de autores já existentes e evitando o surgimento de novas ambiguidades.

Ferramentas podem ter outras aplicações, como no agrupamento de dados de prontuários de pacientes

PADRÕES DE CONECTIVIDADE

A combinação de diferentes metodologias pode trazer resultados mais acurados. Diego Raphael Amancio, pesquisador do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP), desenvolveu um método para solucionar ambiguidades de assinaturas baseado na análise das redes de colaboração dos autores, mas que não se limita a avaliar quem atuou em parceria com quem. Sua estratégia analisa os padrões de conectividade de uma ampla rede de pesquisadores e mostra a situação de cada autor nesse universo. "Utilizando conceitos da teoria de redes complexas, é possível gerar grafos, avaliar a densidade das conexões entre autores e a distância média entre o pesquisador que estou estudando e os demais", explica Amancio, que propôs utilizar tais medidas para caracterizar a produção de um autor e compará-la com a de outro com o mesmo nome, a fim de resolver problemas de ambiguidade. Amancio foi o autor principal de um artigo publicado em 2015 na *Scientometrics* que mostrou a eficiência do uso dessa técnica combinada com a análise de padrões de colaboração já consagrada. Mostrou, em simulações com um conjunto de três bases de dados selecionadas para o estudo, que a capacidade de solucionar ambiguidades dessa solução híbrida chegou a 85%, ante 53% quando apenas a abordagem tradicional era utilizada.

Ao mesmo tempo que amplificou o problema da ambiguidade em referências bibliográficas, o crescimento da produção científica mundial inspirou novas soluções que passaram ao largo dos algoritmos. Em 2012, foi criado um código alfanumérico que serve como identificação única de pesquisadores. Batizado de Orcid (Open Researcher and Contributor ID), o número passou a ser exigido por instituições e agências de fomento e aglutina a produção de cada autor de forma automática (*ver* Pesquisa FAPESP *nº 238*). Mais de 2 milhões de autores já têm sua identificação particular. "Mas nem todos os pesquisadores utilizam esse código e ainda é necessário utilizar métodos empíricos para analisar bibliotecas antigas", diz Alberto Laender. Daniel Figueiredo observa que o conhecimento acumulado no esforço contra a ambiguidade de nomes pode ter outras aplicações. "É possível utilizar as ferramentas em outros contextos", diz. Um deles é o agrupamento de informações de prontuários médicos de um mesmo paciente que foi atendido em hospitais públicos ou postos de saúde diferentes. "Também pensamos em estudar o padrão de uso de nomes com ambiguidade de atores e de cineastas em acervos de filmes, como o Internet Movie Database", informa Figueiredo. ■ **Fabrício Marques**